EOS as a datalake technology

Bird, Campana, Espinal, Girone (CERN-IT/LCG)


processing
clouds
opportunistic cpus, collaborations, supercomputers
storage
glaciers
tapes, nearline
data
torrents
daq, recovery, production
workload
avalanches
analysis, reprocessing, conferences
WLAN and LAN
communication networks
Interlakes, intersites, technologies
lakes
data repos
CERN
WLCG
Worldwide LHC Computing Grid

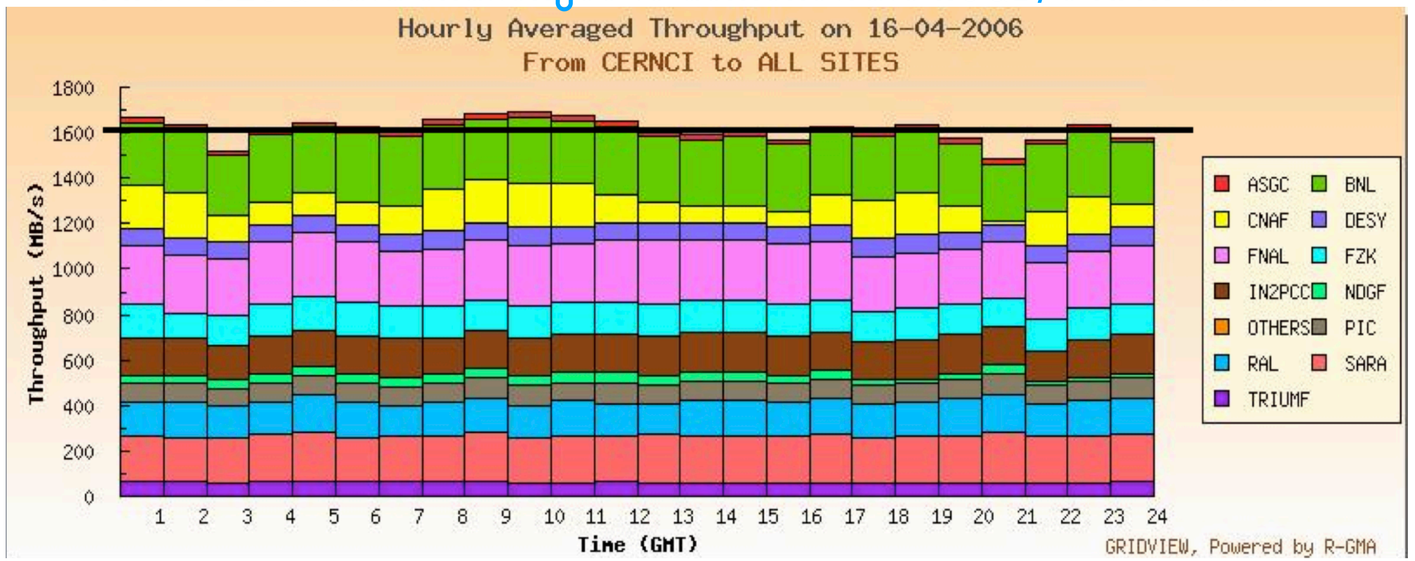
The challenge continues: the goal is unchanged

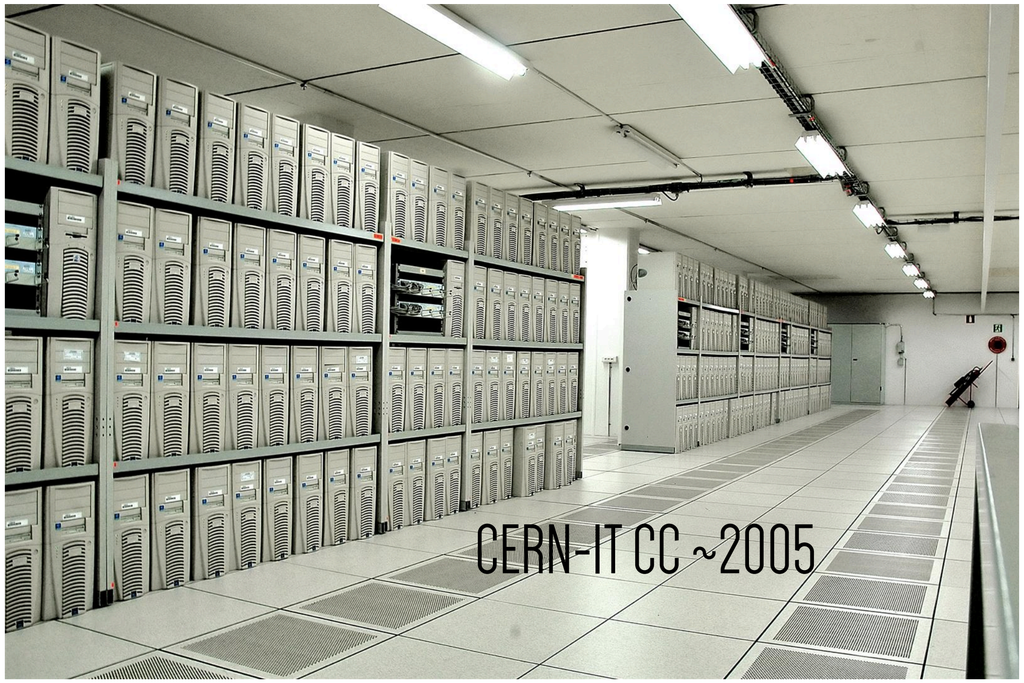
circa 2006

Distributed computing (DC) exploration. WLCG Service Challenges. IPB fit in 8 Racks. Clocks 1.86G/dualcore. 10GE is a dream.

Physical space is an issue (commodity PCs as worker nodes). PUE not yet a figure. Network is scaling.

First WLCG data challenges to demonstrate 1.6GB/s RAW data





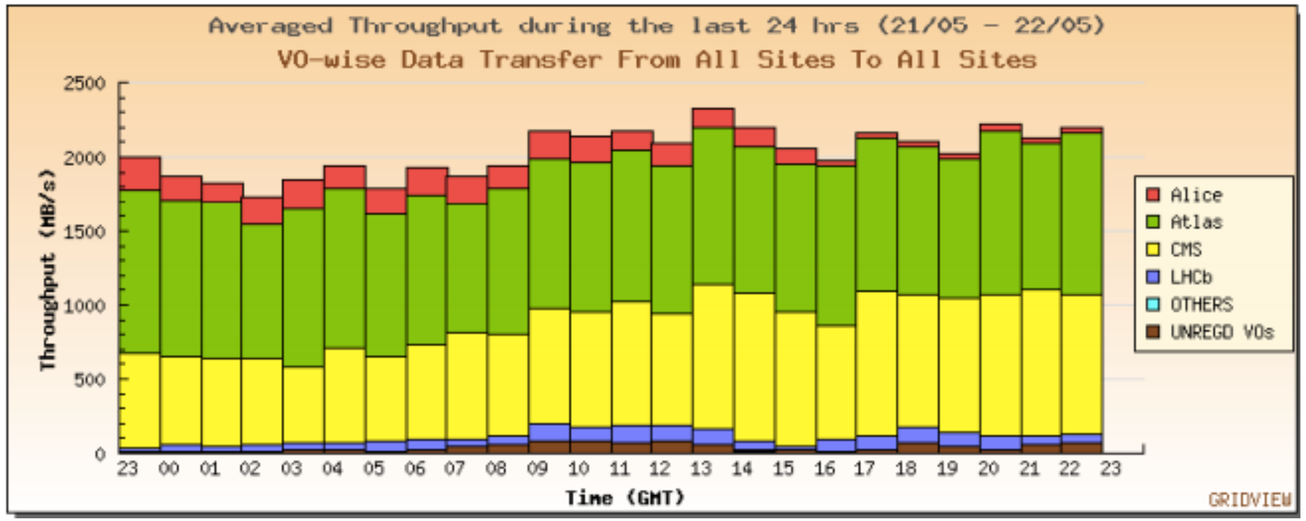
The goal: to provide a computing infrastructure to the experiments and the community to store and analyze data

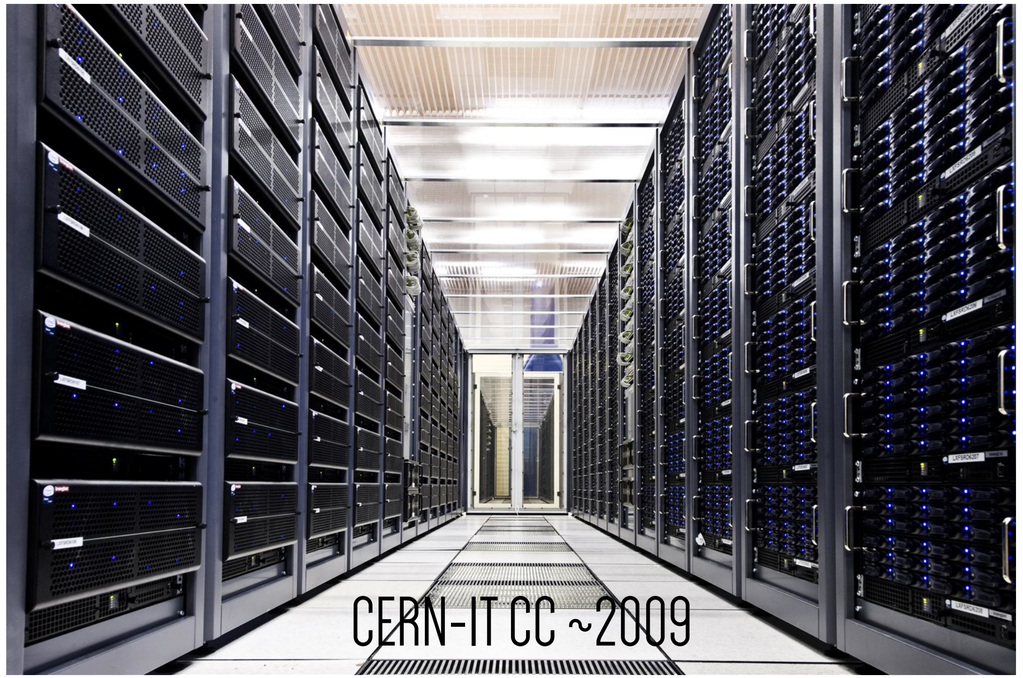
The challenge continues: the goal is unchanged

circa 2009

Phasing Run-I: Distributed Computing is consolidated. IPB fit in 3 Racks. Clocks at 2.67G/quadcore. 10GE is luxury, 100Gbps in the horizon.

Power is an issue. Hot/cold corridors. Compact diskservers, compact-pizza nodes. Heat: PUE is a figure. LANs struggling.





IT-ST

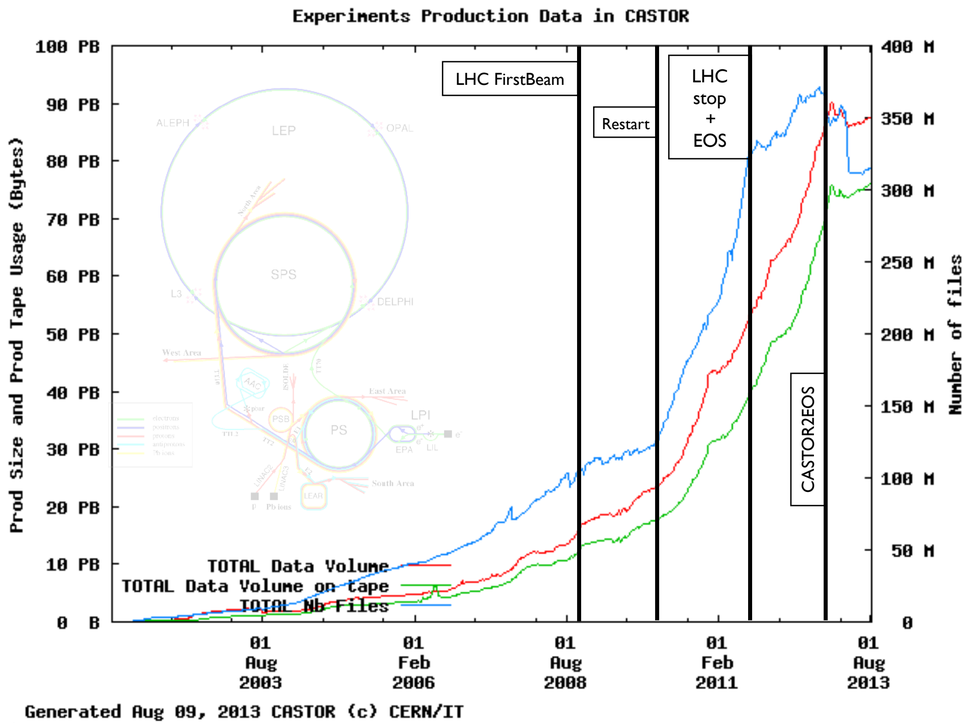
The goal: to provide a computing infrastructure to the experiments and the community to store and analyze data

The challenge continues: the goal is unchanged

circa 2012

Phasing Run-II. DC paradigms shifting. IPB fit in one Rack. Clocks at 2.4G/multicore. 10GE is the standard, some 100Gbps in place.

Power consumption is a figure on tenders. Physical space starts to be freed. Networks upgraded. Actions to tame the PUE.



The goal: to provide a computing infrastructure to the experiments and the community to store and analyze data

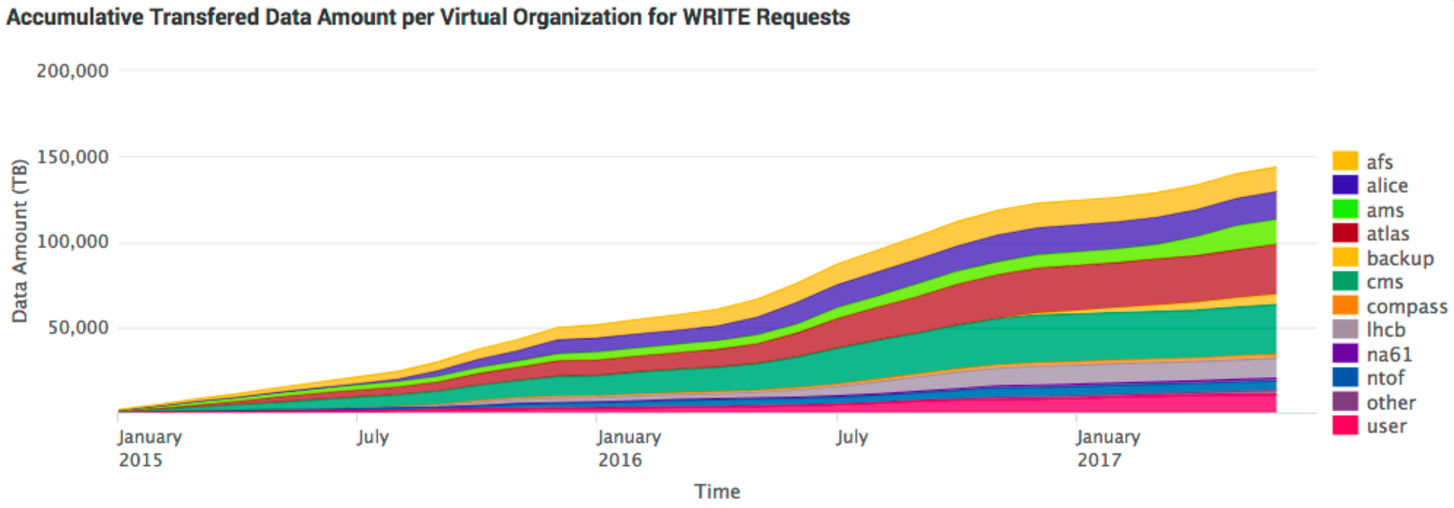
The challenge continues: the goal is unchanged

circa 2017

Ending Run-II. DC re-evaulation. IPB fit in a server (5U). Clocks at 2.4G/multicore. 10GE at the limit, 40GE next standard (?)

CCs getting “empty”. Super racks: +kW, internal cabling. Super-compact servers. Green-IT. $$$ is the limit.

Run2 cumulated data



The goal: to provide a computing infrastructure to the experiments and the community to store and analyze data

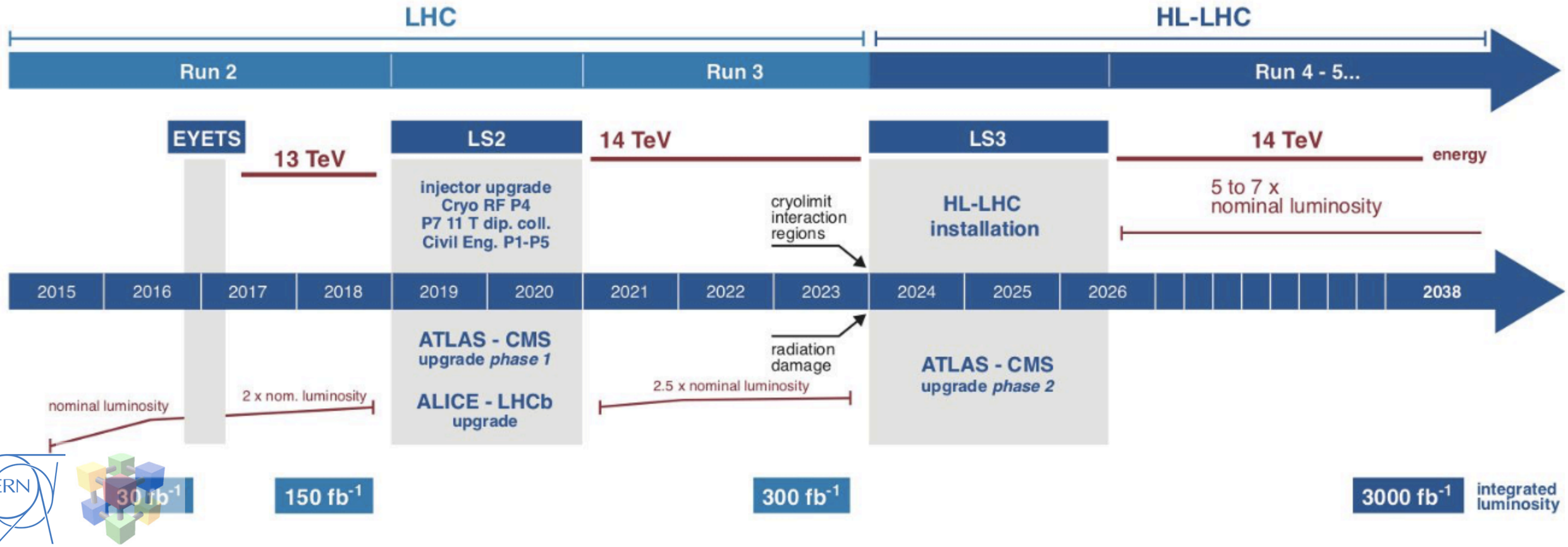
Datalakes motivation: High Luminosity LHC

After long and successfull period (10+ years) WLCG data management and data processing models need to be revised.

Moore/Kryder's law ending and flat budgets does not align well with experiments planning to run at higher trigger rates on a High Luminosity 14TeV p-p accelerator.

Datalakes motivation: High Luminosity LHC

ATLAS and CMS experiments will need one order of magnitude more storage resources than what could be realistically provided by the funding agencies at the same cost of today.



Datalakes motivation: High Luminosity LHC

CERN openlab

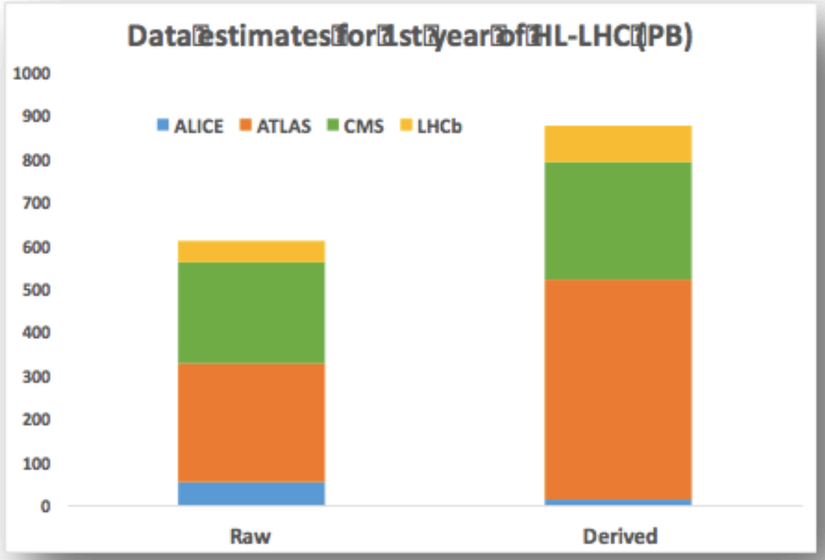
LHC Run3 and Run4 Scale and Challenges

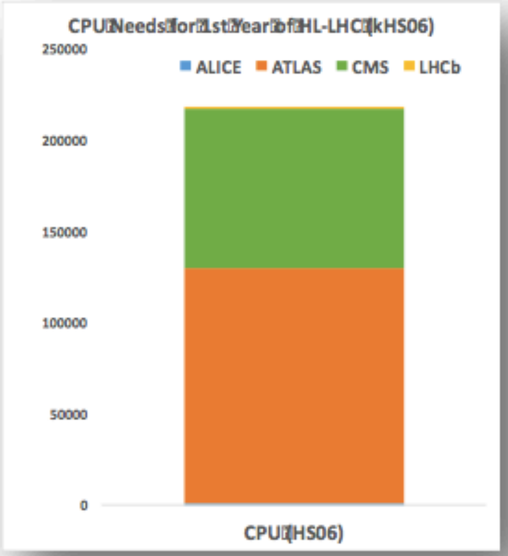
2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 ..... 2030?

First run	LS1	Second run	LS2	Third run	LS3	HL-LHC	FCC?

- Raw data volume for LHC increases exponentially and with it processing and analysis load
- Technology at ~20%/year will bring x6-10 in 10-11 years
- Estimates of **resource needs at HL-LHC x10** above what is realistic to expect from technology with reasonably **constant funding**

Technology revolutions are needed



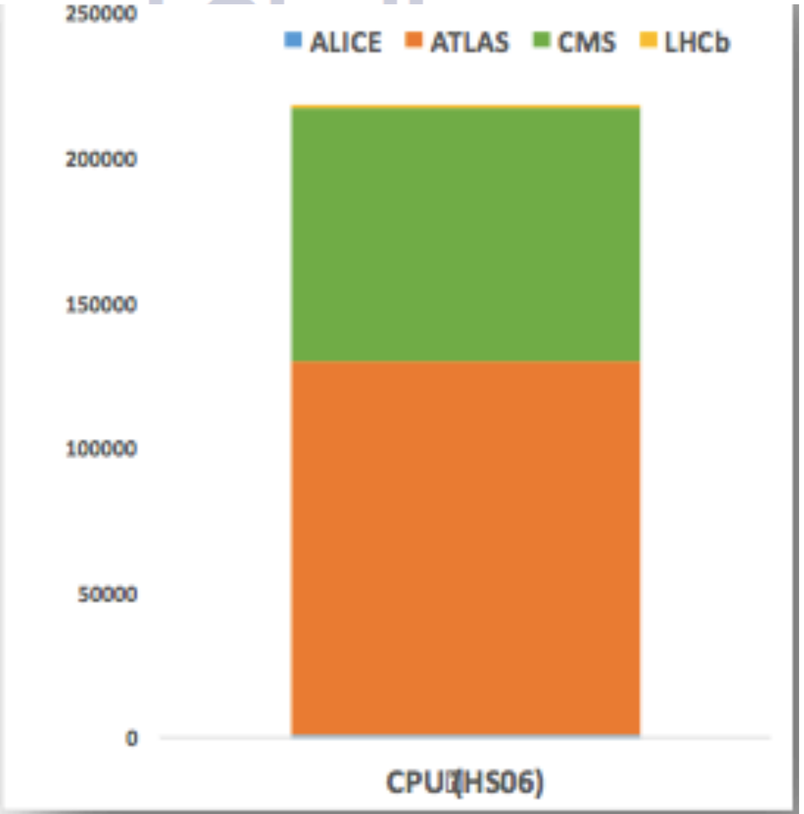


Data:
- Raw 2016: 50 PB → 2027: 600 PB
- Derived (1 copy): 2016: 80 PB → 2027: 900 PB

CPU:
- x60 from 2016

Courtesy of I. Bird

Datalakes motivation: High Luminosity LHC



Data:

- Raw 2016: 50 PB → 2027: 600 PB
- Derived (1 copy): 2016: 80 PB → 2027: 900 PB

CPU:

- x60 from 2016

Datalakes: a distributed storage evolution

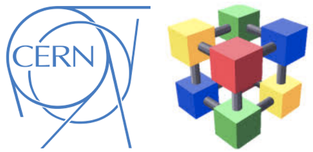
Datalakes are an extension of storage consolidation where geographically distributed storage centers (potentially deploying different storage technologies) are operated and accessed as a single entity.

The goals

Optimize storage usage to lower the cost of stored data

- The enabling technology should be based on: geo-awareness, storage tiering and automated file workflows fostered by fa(s)t networks.

Trigger a revision of the current computing models to re-evaluate data management, data access and data consolidation practices.

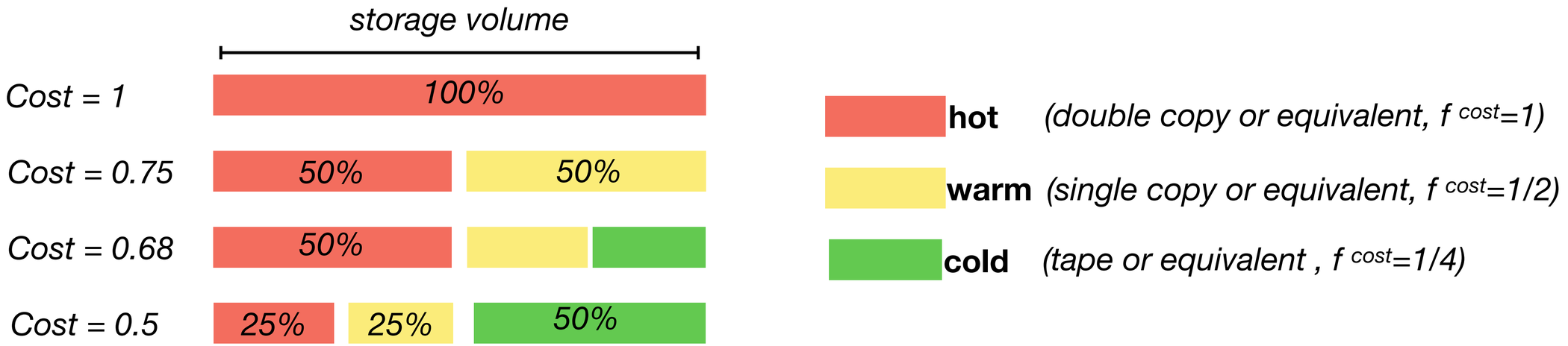


Leverage data costs by QoS

Treating the entire data repository equally does not scale. ~25% of the entire dataset is actively used (and no predictions possible),the reamining percentage can be 'treated' to reduce costs. Needless to say this economic savings impacts on reliability or/and performance.

Hierarchical storage concept driven by cost. Storage tiering through file workflows.

What experiments would chose? N PB of data at 10^{-5} reliablity or 2xN PB at 10^{-4} reliablity?



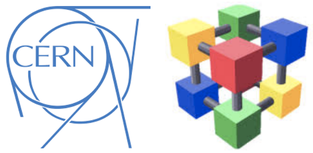
EOS evaluation as a Storage System candidate - why?

EOS is the main storage system for LHC and SME at CERN (250+PB). Conceived, made, maintained, evolved and shaped at CERN for the physics (and not only physics anymore) community.

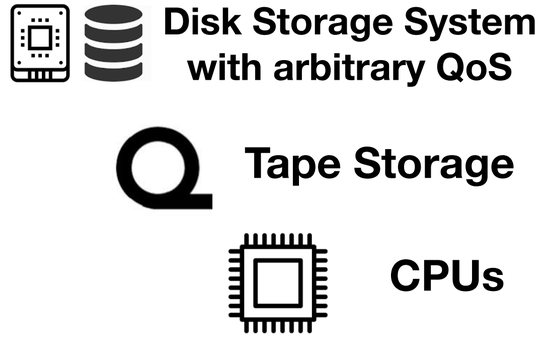
EOS allows to define different QoS levels by file or directory, allows file transitioning and can handle different storage technologies as filesystems (SSDs, Enterprise HDDs, Consumer HDDs and Kinetic Drives).

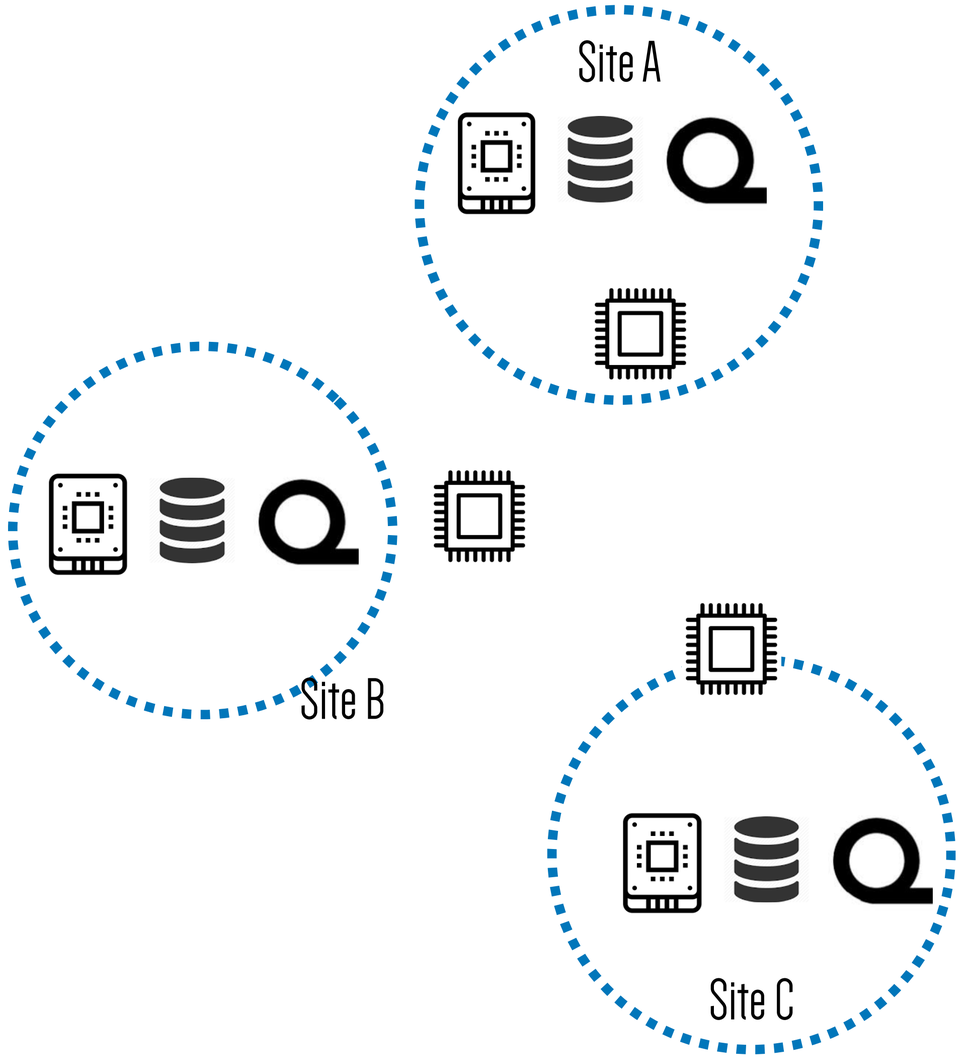
Scalability has been largely demonstrated and new namespace (RocksDB) breaks the in-memory size constrains opening the multi-B files possibility.

Geo-awareness is a built-in capacity in production (GE/HU) and allows different granularity levels: country, room, rack or server.

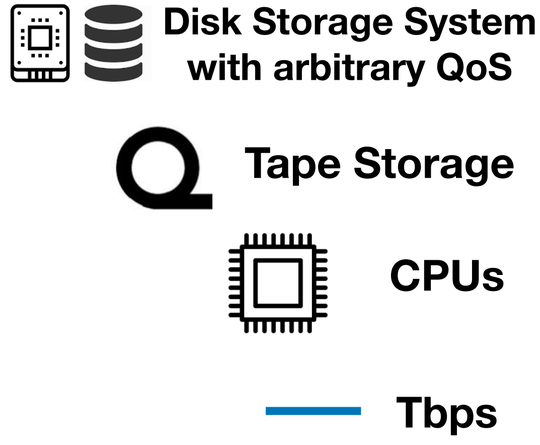


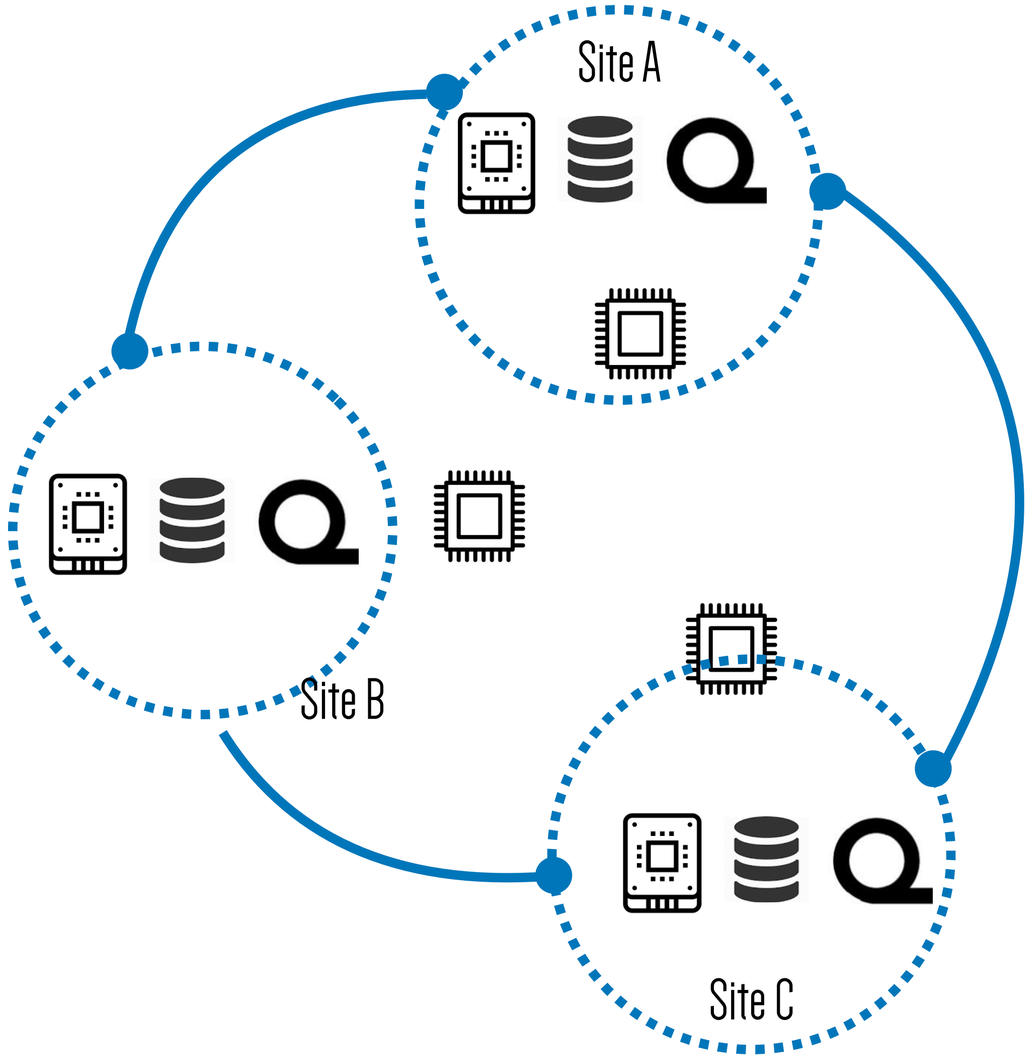
Datalake example - data flow



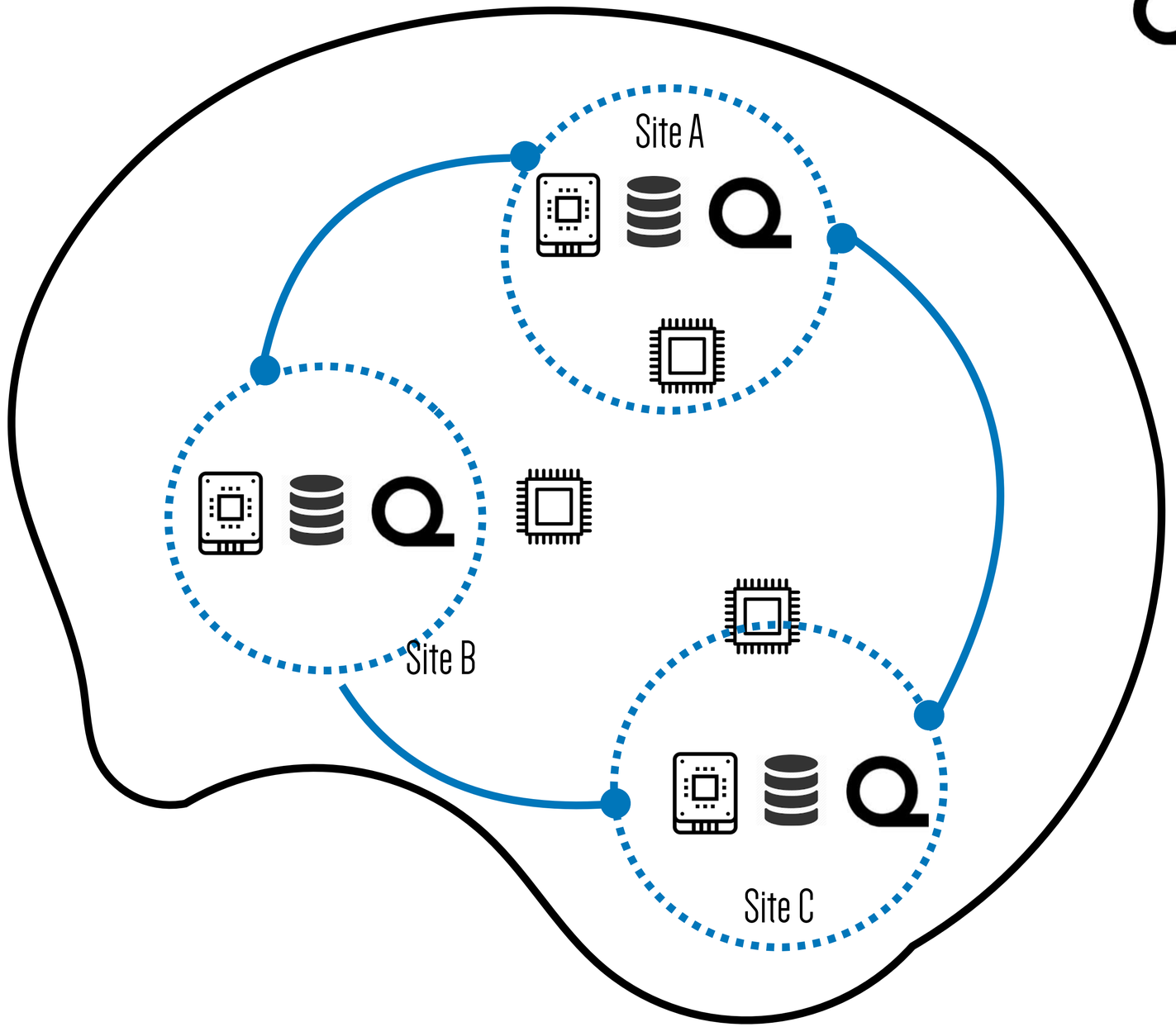


Datalake example - data flow





Datalake example - data flow



Disk Storage System with arbitrary QoS

Tape Storage

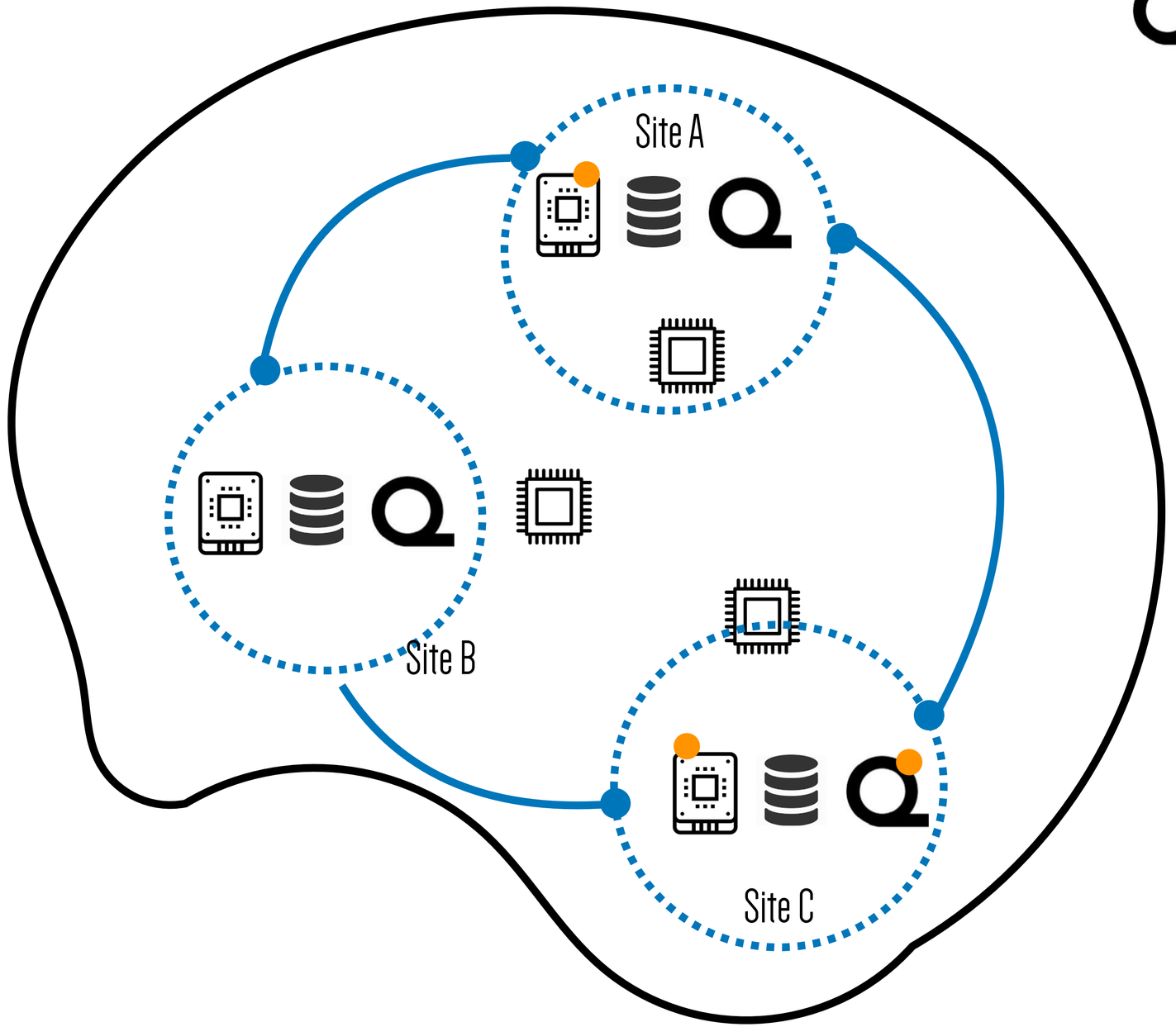
CPUs

Tbps

Datalake example - data flow

File placement by QoS

- Hot custodial file (2 fast copies+archive)



Disk Storage System with arbitrary QoS

Tape Storage

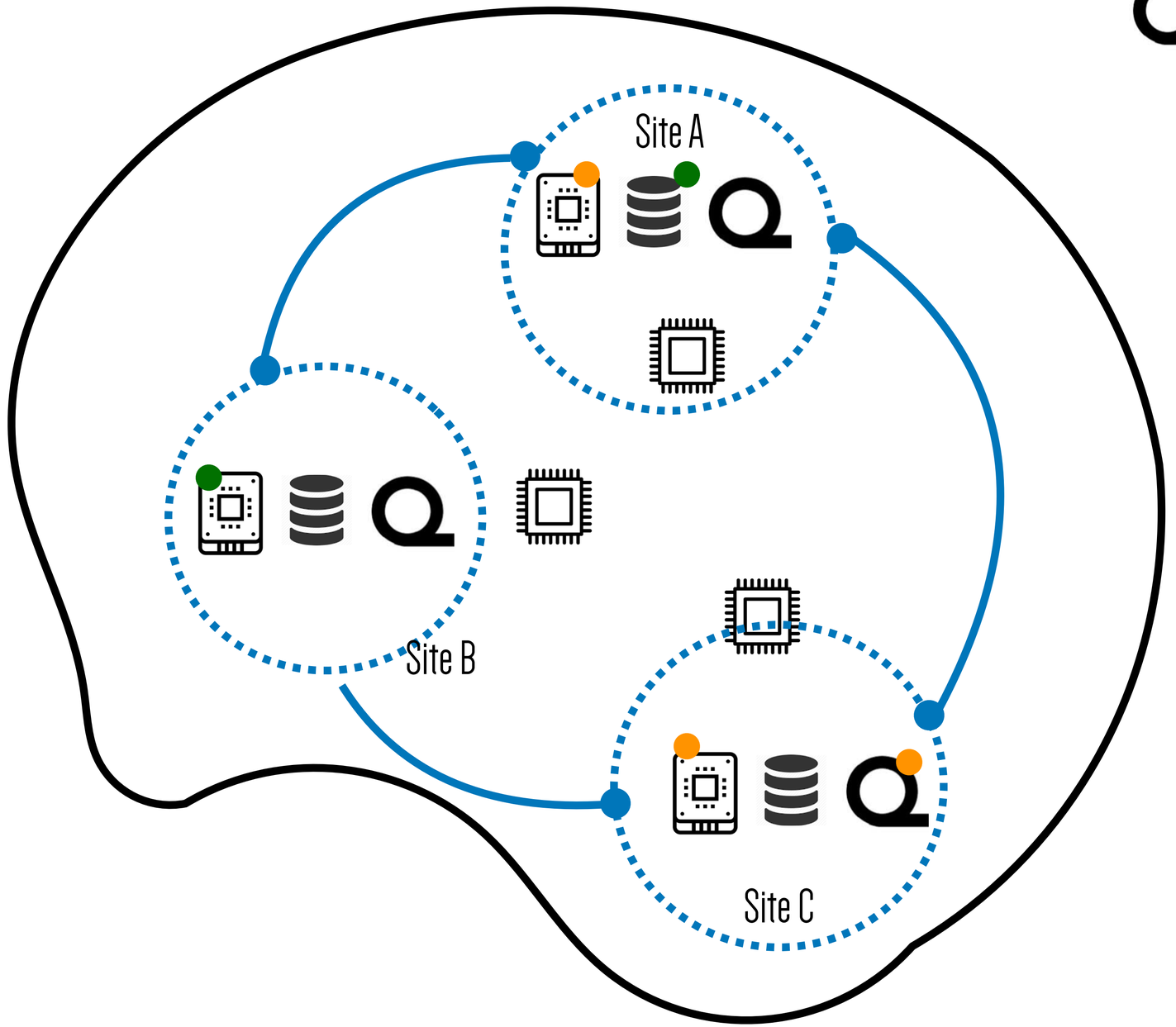
CPUs

Tbps

Datalake example - data flow

File placement by QoS

- Hot custodial file (2 fast copies+archive)
- Hot ephemeral file (2 fast copies)



Disk Storage System with arbitrary QoS

Tape Storage

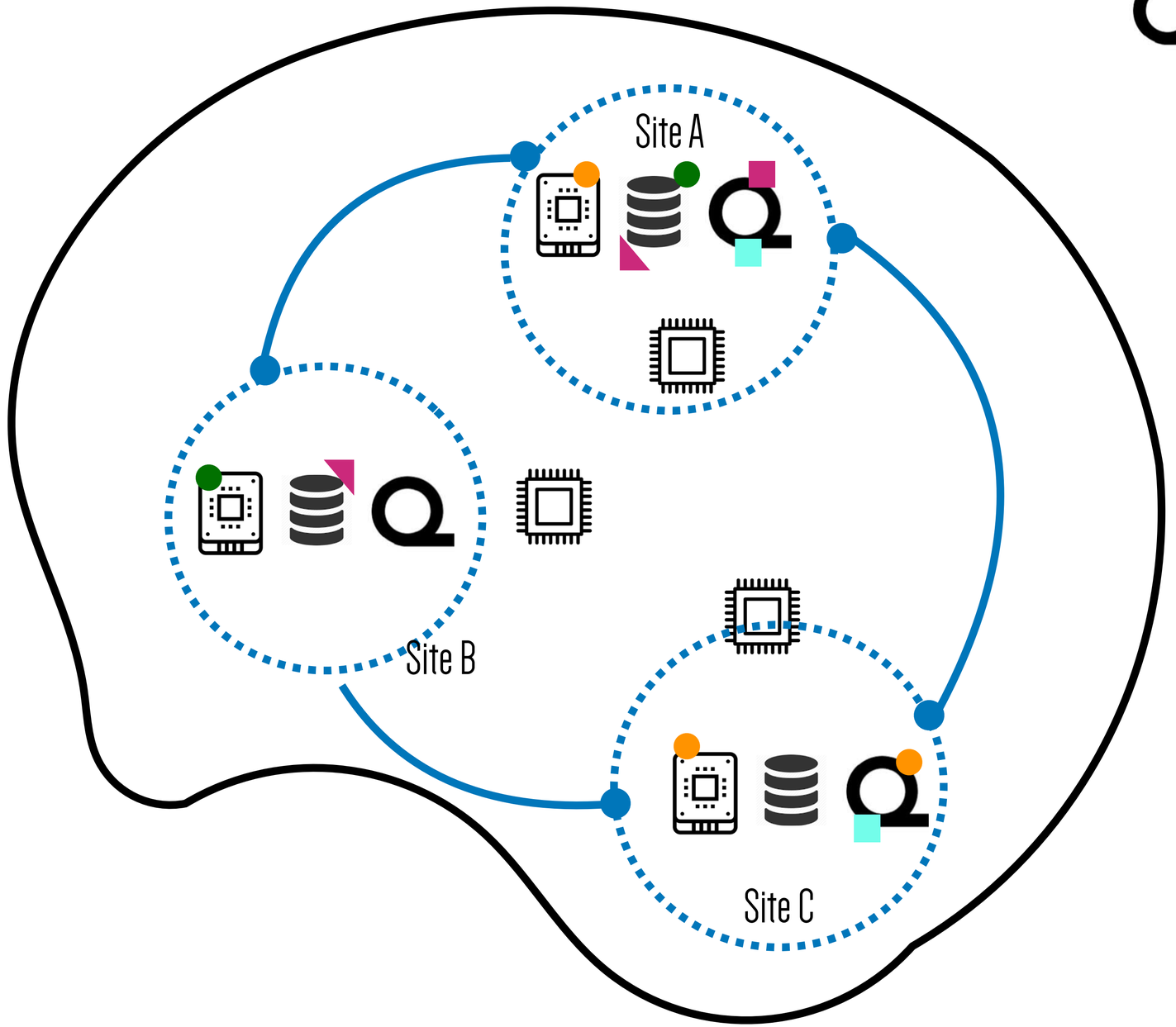
CPUs

Tbps

Datalake example - data flow

File placement by QoS

- Hot custodial file (2 fast copies+archive)
- Hot ephemeral file (2 fast copies)
- Warm ephemeral file ("Rain")



Disk Storage System with arbitrary QoS

Tape Storage

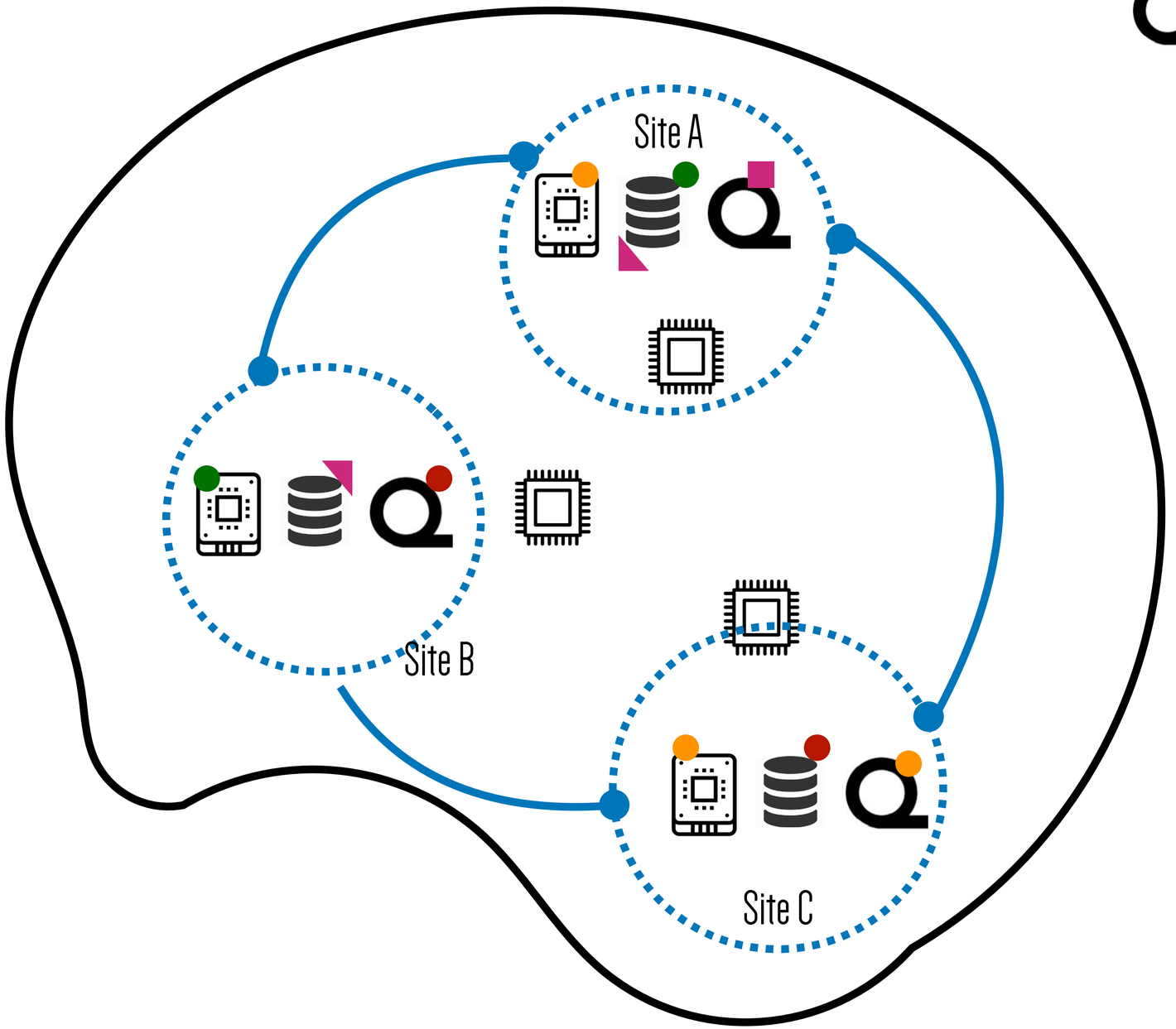
CPUs

Tbps

Datalake example - data flow

File placement by QoS

- Hot custodial file (2 fast copies+archive)
- Hot ephemeral file (2 fast copies)
- Warm ephemeral file (“Rain”)
- Warm file (disk copy+archive)



Disk Storage System with arbitrary QoS

Tape Storage

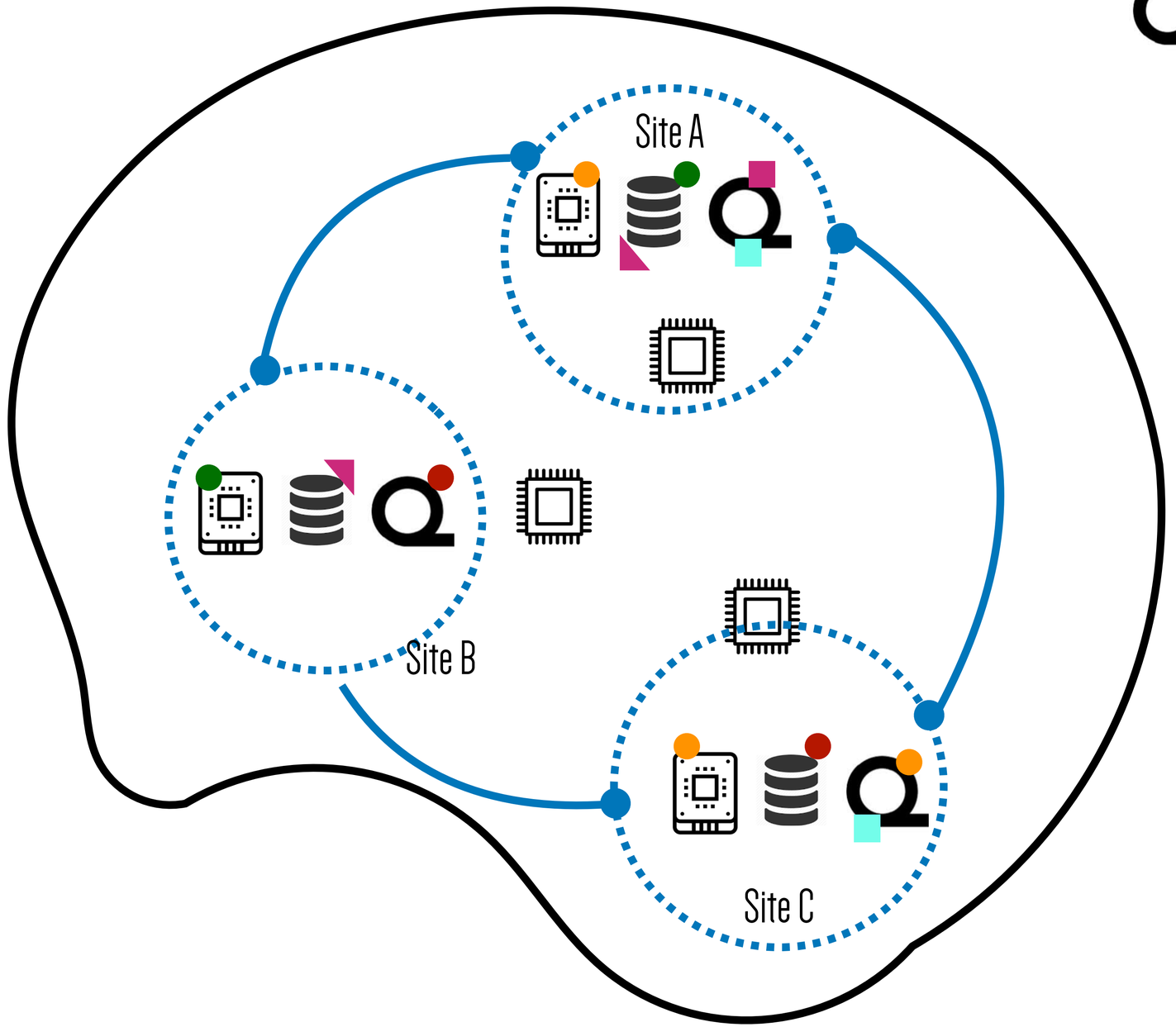
CPUs

Tbps

Datalake example - data flow

File placement by QoS

- Hot custodial file (2 fast copies+archive)
- Hot ephemeral file (2 fast copies)
- Warm ephemeral file (“Rain”)
- Warm file (disk copy+archive)
- Cold file (archive)



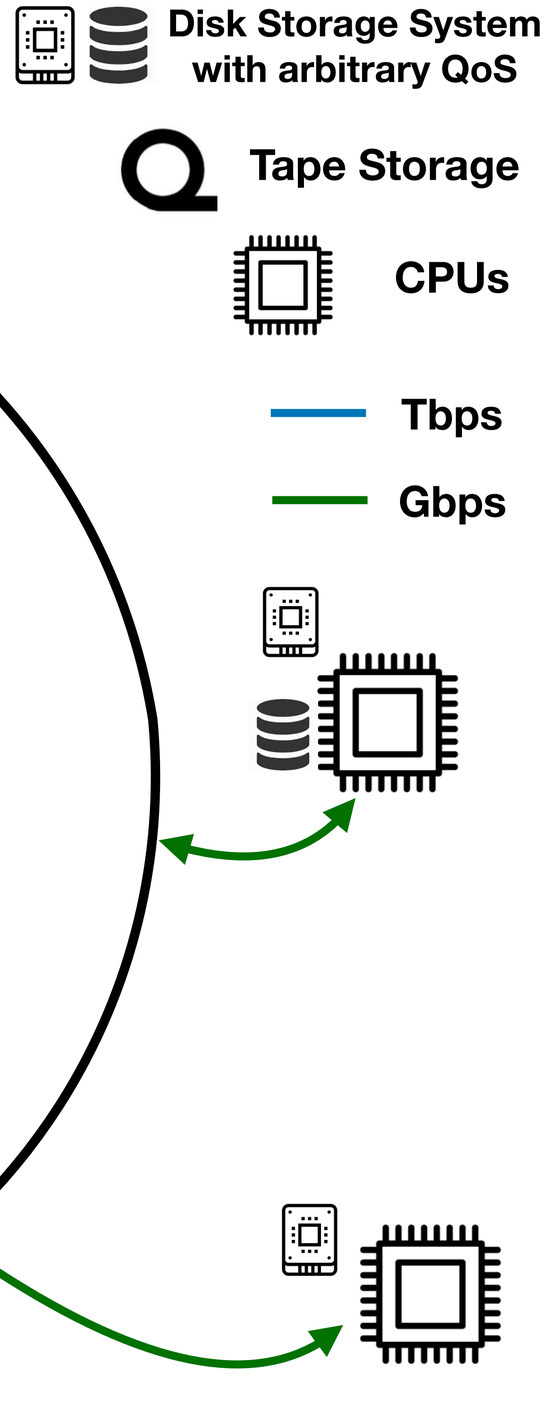
Disk Storage System with arbitrary QoS

Tape Storage

CPUs

Tbps

Datalake example - data flow



File placement by QoS

- Hot custodial file (2 fast copies+archive)
- Hot ephemeral file (2 fast copies)
- Warm ephemeral file ("Rain")
- Warm file (disk copy+archive)
- Cold file (archive)

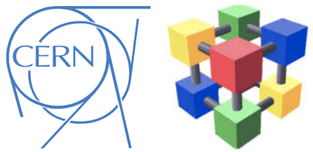
Summary and outlook

Change of paradigm needed to fulfill HL-LHC data storage capacity requirements by 2020.

Datalakes (WLCG meaning) observed as a good candidate to start evolving the long successfull current model.

The keys for success to build a datalake-based model strongly depends on the capability to leverage the storage at the sites in a centralized manner. Experiments drive the required QoS statically and dinamically. Full dataset perceived as a whole entity.

R+D task within WLCG to build up a datalake demonstrator with EOS as enabling technology. Activity coordinated by CERN-WLCG with the participation of selected HEP labs.



Thanks